Predicting Wages and Recruitment Rates in Swedish Regions and Municipalities

Youheng Lü
Department Microdata Analysis
Dalarna University
Borlänge, Sweden
h21youlu@du.se

***Abstract*—Finding the right place to look for a well-paying job can take a lot of time that would be better suited elsewhere. While other works have analysed general trends in wages and recruitment, we have not found anything that analyses the differences within regions of a nation. This paper tackles this problem by creating models for the predicting wages and recruitment rates in Swedish Regions and Municipalities. It establishes a Linear Regression prediction model. Using 5-fold Cross-Validation to compare the accuracy against Tree-based models, this paper provides a baseline model that future work can build on. The key takeaways are that a Linear Regression model is accurate at predicting Wages in Sweden with a R^2 of 0.893. For predicting Recruitment rakes we tried two different approaches of how Quarters within our Linear Regression model. Both models provided low accuracy but the comparison to Decisions Trees showed it might not be the problem of the model but the volatility of the data. Further research might require a different dataset to give an accurate recruitment rate prediction. Still, we found out that the region Stockholm has higher recruitment rates than any other region in Sweden.**

***Index Terms*—linear regression, wage prediction, recruitment rate, decision tree, random forest**

I. Introduction

Predicting wages and recruitment rate is an active research topic [1] [2] [3] [4] [5]. This question is especially important for future University graduates are looking for opportunities to find well-paying employment in Sweden [6]. Research shows that workers who set a lower reservation wage if they are more impatient. [7], which could suggest that Job-Search is so tedious, that workers are willing to accept a lower wage. To assist in this process, we want to investigate how to build a model which can predict wages and recruitment rates based on region and municipality. We attempt to answer the following two research questions:

A. Research Questions

1) How would we create a model that accurately predicts wages for Sweden's municipalities?
2) How would we create a model that accurately predicts recruitment rates for Sweden's regions?

We organize the paper as follows: Section II provides a general literature review of relevant topics. Section III provides a general overview of the dataset and the methods used to build and evaluate the prediction model. In Section IV we show the results of our analysis which we discuss in Section V. We finish this paper with Section VI which provides a Conclusion and ideas for future work.

II. Literature review

On the topic of predicting wages one paper uses techniques for imputing missing pay data and forecasting salaries over a 6-year period. [5]. The paper developed model specifications based on the link between salaries and other macroeconomic and labour market factors, and then employed pseudo out-of-sample testing to pick the top performing models by nation. This technique allowed to produce of a balanced panel data collection for 112 nations throughout the period 1995-2019. Another work focused specifically on Sweden. The paper did analysis of the relations between private and public sector wages. It discovered two long-run correlations between central government, local government, and private sector salaries in Sweden using a maximum

likelihood cointegration technique [8]. More recent work [10] has used machine learning techniques like XGBoost and Deep leaning to achieve accurate Wage prediction of the Current Populations survey in the United States. Still, they were not able to outperform the empirical Mincer earning function [11] that predict wages based on a function of w_0 baseline, s years of schooling and x potential labour market experience. As well as parameters ρ, β_1 β_2:

$$\ln w = f(s, x) = \ln w_0 + \rho s + \beta_1 x + \beta_2 x^2 \quad (1)$$

Both the machine learning competition and the mincer earning function assume knowledge of other know features of a specific person to be able to predict the wages. Other work predicted the minimum salary with Wavelet analysis and adaption methods in EU-Member states [3]. Another work [12] analysed the general wage structure between 1963 and 1989, analysing the wages of college graduates with *x* years of schooling. The goal was to detecting patterns like a linear trend and explaining variation through business shocks and deficits in global trade.

Work related to employment prediction exists in the following areas: One paper analysed the effectiveness of a model based on job search-related web queries in predicting quarterly unemployment rates in small samples [2]. Work has also been done on the analysis of how Employment is distributed in Sweden over the Years, showing that local government employment rose from 1959 - 1993 [4]. Another paper focused a probabilistic analysis of the likelihood on finding a job [1] given the length of unemployment. The result show that people with longer unemployment have lower chances of finding employment. One paper [13] which supports the assumption of our research found that local labour shortages lower unemployment rate for disadvantaged youths.

In the end we did not find any work that builds a model that helps jobseekers predict the current environment to find the best opportunity for a job which is where our paper aims to fill this gap.

III. Material and Methods

For the wage prediction we use the dataset "Average monthly salary in the municipalities by municipality and sex, year 2007 - 2021" [14]. This dataset contains 12264 rows and 4 columns. The average monthly salary refers to employees aged 18-64 who can report salaries to year 2013 and refers to employees aged 18-66 for whom we can report salaries from year 2014. If an activity could not be assigned to a specific salary it was ascribed to *3000 Local federations*. We put a detailed description in Table I.

For the prediction of recruitment rate, we use the dataset "Recruitment and vacancy rate, Business by region NUTS2, Quarter 2015K2 - 2021K4" [15]. The dataset also contains observations for vacancy rate but since we are only interested in recruitment rate, we drop them. SCB describes Recruitment rate in this dataset as proportion of vacancies divided by the number of employees. The dataset then contains 243 rows and 3 columns. We put a detailed description in Table II.

A. Model

Andreas Jokobsson has provided a general overview of Time-Series modelling [9]. His book describes Multiple methods how to model time-Series data. This paper uses his methodology described in Chapter 2.2.4 on Linear Projections on normal distributed vectors to make Predictions of the future. We want to see whether a Linear regression model can fit, since it is easy to scale for time series data, where only the year will change, and all the other variables stay the same. To see whether that applies we first plot the data to gain first insights. After gaining those insights, we modify the data accordingly for the model to be applicable. On

TABLE I
DESCRIPTION OF AVERAGE MONTHLY SALARY DATASET

Feature	**Type**	**values**
sex	categorical	*men*, *women*, *total* (men + women)
region	categorical	290 unique municipalities
year	numerical	2007 - 2020
average wage	numerical	avg. monthly wage in SEK

TABLE II
DESCRIPTION OF RECRUITMENT RATE DATASET

Feature	**Type**	**values**
region	categorical	8 unique regions + 1 aggregated
quarter	string	2015K2 - 2021K4
recruitment rate	numerical	Recr. rate for Business sector

this data we create a Linear Regression model. To evaluate the model, we use the *skikit-learn* library: Since we are dealing with a regression problem, we cannot use the metrics like error and accuracy which are only applicable for Classification. We use 5-fold Cross Validation that calculates the average R^2 for 5 different validations sets. The R^2 is a metric that shows how close the model predictions are to the real data. We can assume that a good model needs to have a high R^2 score: The best possible score is 1.0 but the scores can be negative. The scikit-learn Website [16] provides a more detailed description. Since we are using validation sets, we can assume that R^2 is an accurate scoring mechanism: Even if it has a high R^2 we can be quite confident that we are not over-fitting. We compare the R^2 of Linear Regression to a Decision Tree and a Random-Forest with Hyper parameter Tuning. The R^2 of the Decision tree provides a baseline how good the model can be since it does not have any assumptions about the distribution of the data. The Random-Forest should perform better than the Decision tree, so we also fit that. We only use the tree-based models for evaluation since we expect them to perform badly for predicting Time-Series data. Given the output is based on decision nodes, all data of the future will fall through the same node (e.g. years >2020) which means that there are no changes in the prediction.

1) Recruitment rate prediction: We approach Recruitment rate prediction can in multiple ways since the data gives us not only the year but also quarter. We first split the column *Quarter* into two columns *Year* and *Quarter* to be able to work on those numerically. We then create a new variable called *Time* that has the formula:

$$Time = Year + (Quarter - 1) \cdot 0.25 \quad (2)$$

which translates each quarter to a $\frac{1}{4}$ of a year. With those two predictors we propose two linear models to predict the response variable Y (*Recruitment rate*) that can take two forms:

$$Y = a_{Intercept} + a_1 \cdot Year + a_{quarter} + a_{region} \quad (3)$$

This first model assumes that there are quarterly differences, so we use the quarters as a categorical variable or Form 2:

$$Y = a_{Intercept} + a_1 \cdot time + a_{region} \quad (4)$$

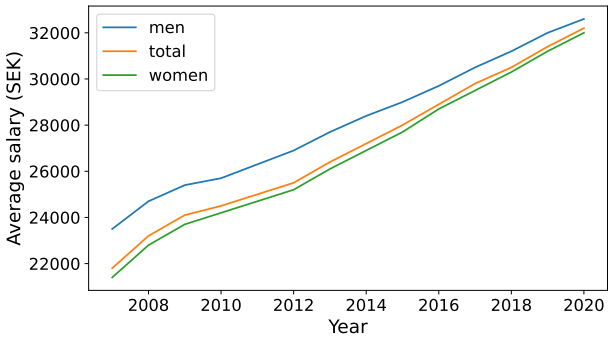

Fig. 1. Average monthly salary in Sweden, split by *men*, *women* and *total*, which is the average monthly salary of men and women combined.

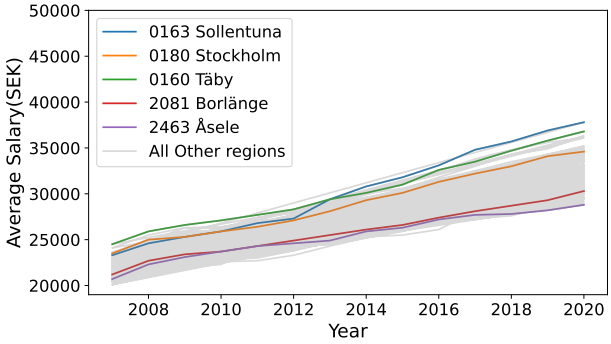

Fig. 2. Average monthly salary in selected regions. Sollentuna and Stockholm start at a similar wage in 2007 but reach different endpoints, Täby has highest wage in 2007 but is low in 2020. We choose Borlänge since the audience should be familiar municipality. The grey lines show the wages for all other regions.

The second model minimizes variance by cutting down on features: I use the variable *time* instead. This model is supposed to perform well if a general linear in the Wages data exists. It also has the advantage of being the simpler model, which given our small dataset might be a better fit.

IV. Results

A. Wages prediction (Q1)

We first plot the data to find interesting patterns we want to use later for modelling. A first look at the average wage in all of Sweden Fig. 1 shows that wages have been increasing at a very steady rate. We can also see that the wage gap between men and women shrank between the years. For sake of simplicity, since the data could affect our Cross-Validation, we will drop the *sex* column and only

TABLE III
MODEL COMPARISON WAGE PREDICTION

Model	R^2
Linear Regression	0.893
Decision Tree	0.904
Random Forest	0.892
Tuned Random Forest	0.902

look at the total wages. Since we are only interested in the municipalities, we also dropped the datapoints referencing the national level *00-Sweden* and the *3000-Local Federation*.

In Figure 2 we can see that there is a general linear upwards trend in the municipalities, which shows that Linear Regression model should fit the data well. On the other hand, we can also see that different municipalities can have different trajectories in the case of Sollentuna versus Stockholm, where they both start at a similar wage but reach a 4000 SEK difference in 2020.

We applied One-Hot-encoding on the column *region* so our model can work with this categorical data. In TABLE III Linear Regression has a R^2 of 0.893 out of a maximum of 1. It has comparable results to a Decision Tree and Random Forest model which suggest that the model is already accurate and suitable for prediction.

The result of the Linear Regression model for wage prediction is shown in Figure 3. As we can see the model predicts based not only on the average wage from 2020 but from the whole data range from 2007 to 2020 of the municipality. The strength of the approach is that it takes the historical development into account and will predict accurately in the future. One weakness could be that the values for 2021, 2022 might be inaccurate, which we see in case of Sollentuna (blue), where the prediction shows a jump in Wages in 2021 for some regions that stems from taking the average over the years as the predictor. To give general overview how the predictions look like we show the top 4 regions with the highest wage in 2027 in TABLE IV. As a comparison we also show the predictions of the Decision tree in Figure 4. We can see that it just results in the naive model where the predictions of 2020 are repeated. This is due to the nature of the decision tree that it performs well for values that lie within the training set but worse for values outside the training set.

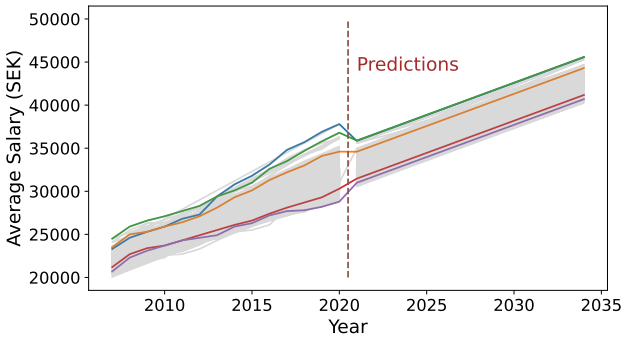

Fig. 3. Prediction with **Linear Regression** of total Wages for all the municipalities from 2021 - 2035. The same regions like in Figure 2 have been chosen, so the legend was removed to prevent distraction from the data.

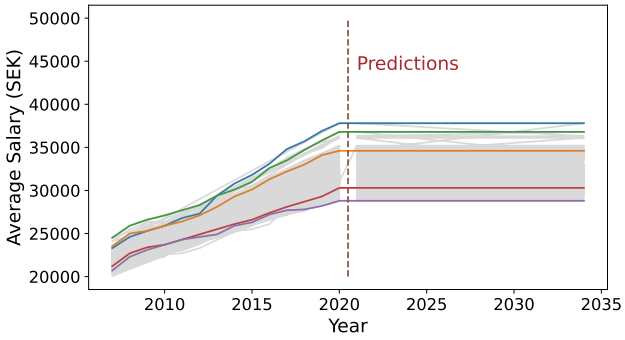

Fig. 4. Prediction with **Decision Tree** of total Wages for all the municipalities. The same regions like in Figure 2 have been chosen. We can see that the prediction stays the same after 2021.

B. Recruitment rate prediction (Q2)

We compared the different Quarters as well to see if there was a pattern specific to quarters. During the analysis we find out that Q1 and Q2 have higher recruitment rate than Q3 and Q4. We notice the fall in recruitment rate in 2020 during the COVID-

TABLE IV
TOP 4 MUNICIPALITIES WITH THE HIGHEST WAGES IN 2027

Region	**Year**	**Average monthly salary**
0160 Täby	2027	40368.779
0163 Sollentuna	2027	40333.065
0182 Nacka	2027	40083.065
0114 Upplands Väsby	2027	39461.636

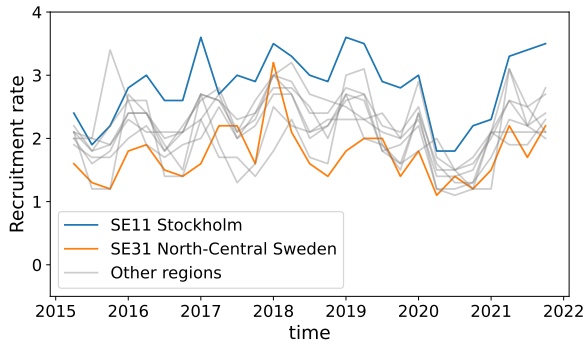

Fig. 5. Recruitment rate in the 9 unique regions of Sweden by total time. We highlighted the Regions SE11 Stockholm and SE31 North-Central Sweden since they should be most familiar to the target audience of this paper.

TABLE V
COMPARISON OF RECRUITMENT RATE PREDICTION MODEL

Model	R^2
Linear Regression Equation 3	-0.029
Decision Tree Equation 3	-0.199
Linear Regression Equation 4	-0.176
Decision Tree Equation 4	0.001

19 Pandemic as well as the subsequent rise in 2021.To see if there is a general trend, we plot the development of Recruitment rates over time. In Figure 5 we can see that there are times where most regions recruit more and times where most regions recruit less but is hard to pinpoint specific trends over the quarters as well. We can also see a clear difference between the region SE11 Stockholm and SE31 North-Central Sweden. This suggests that region vary in the height of the employment rate. We cannot find specific patterns that are useful for modelling given that there is also high variance in the data. Adding the fact that our dataset is also small, we can expect models to perform poorly.

We fit the models based on Equation 3 and Equation 4 with *recruitment rate* as the response variable. The results are show in TABLE V. As expected, we can see that all models perform badly, with R^2s of zero or below and high Variance. Decision-Trees do not perform good either, we therefore conclude that the current data is not enough to make accurate predictions.

V. DISCUSSION

We observed a general linear upwards trend for wage data with the wage difference between men and women becoming smaller over the years. Our Linear Regression model performs well with a high R^2. It even has similar performance relative to Decision Tree and Random Forest which we used for evaluation. Using this model for prediction, we can see that it captures the general trend of the data well. One caveat is the prediction in the first few years, where there are big jumps between 2020 and 2021. The model should therefore perform best for medium-term predictions around 5 years in the future. Further work can apply Moving average smoothing to improve this model so which smooths predictions for short-term and might be more accurate.

With our recruitment rate dataset, we found that the data shows no clear pattern and has high variance. We tried two approaches: Splitting the data by quarters and modelling just with the *time* feature. Both approaches worked poorly, with low R^2s and high Variance in the model. The models created are therefore not suited for prediction. Still, we made two general observations that should be helpful: Firstly, Q1 and Q2 have higher recruitment rates than Q3 and Q4. Second, SE11 Stockholm has comparatively higher recruitment rates than the rest of Sweden. The question remains how to create a better model that accurately predicts recruitment rates. We will provide our thoughts in Section VI about potential future work.

VI. CONCLUSION

We wanted to work on research that helps jobseekers and especially future University graduates in Sweden find well-paying employment. While the topic of wage and employment analysis has been explored before, we did not find predictions models for Swedish regions and municipalities. Therefore, we attempted to create models that accurately predict wages and recruitment rates in Swedish regions and municipalities. We use two datasets from Statistics Sweden [14] [15], one for Wages and one for recruitment rate to build our models. The models we build were Linear regressions on the two predictors time and region with the theoretical basis on basic time series modelling [9]. For recruitment rates

we tried two different approaches for dealing with quarterly data. To evaluate our Linear regression models, we compared against tree-based models using the R^2 which we calculated via 5-fold cross validation. The idea is that tree-based models like Decision Trees and Random Forest with the idea that those models perform well in general and give an accurate baseline of the possible model performance. The result was that wages had a stable linear trend during the last 13 years. Therefore, our Linear Regression model performed well by having a high R^2 of 0.893 which had comparable scores to Decision Tree and Random Forest. In contrast, the recruitment rates vary over time, and we could not observe a general trend. Both our approaches in dealing with quarterly data performed poorly: Linear Regression and Decision Tree had $R^2 < 0.01$ in cross-validation. Still, we found the interesting result that Stockholm had higher recruitment rate than the other regions.

A. *Future work*

To improve on this model for predicting recruitment rates, we propose the following steps: The first step would be to find a more comprehensive dataset, for example one that also includes vacancy rate of the regions. Another way would be to include a more comprehensive dataset of municipalities that contains more than the nine regions of the current dataset. With that, researchers could consider other modelling approaches that take the trend and seasonality of the data into account. Future work should also consider Moving average smoothing.

REFERENCES

[1] Shimer, Robert. "The probability of finding a job." American Economic Review 98.2 (2008): 268-73.

[2] Francesco, D'Amuri. "Predicting unemployment in short samples with internet job search query data." (2009).

[3] Hadas-Dyduch, Monika. Model-spatial approach to prediction of minimum wage. No. 29/2016. Institute of Economic Research Working Papers, 2016.

[4] Rosen, Sherwin. "Public Employment and the Welfare State in Sweden." Journal of Economic Literature, vol. 34, no. 2, 1996, pp. 729–40. JSTOR, http://www.jstor.org/stable/2729220. Accessed 20 May 2022.

[5] Ernst, Ekkehard, et al. "Predicting Wages." (2016).

[6] https://studyinsweden.se/moving-to-sweden/work-internships/

[7] DellaVigna, Stefano, and M. Daniele Paserman. "Job search and impatience." Journal of Labor Economics 23.3 (2005): 527-588.

[8] Jacobson, T., Ohlsson, H. Long-run relations between private and public sector wages in Sweden. Empirical Economics 19, 343–360 (1994). https://doi.org/10.1007/BF01205942 https://studyinsweden.se/moving-to-sweden/work-internships/

[9] Jakobsson, Andreas. An introduction to time series modeling. Studentlitteratur, 2019.

[10] Ghei, Dhananjay, and Sang Min Lee. "Annual Wage Prediction: Machine Learning Competition." (2020).

[11] Mincer, Jacob. "Investment in human capital and personal income distribution." Journal of political economy 66.4 (1958): 281-302.

[12] Kevin M. Murphy, Finis Welch, The Structure of Wages, The Quarterly Journal of Economics, Volume 107, Issue 1, February 1992, Pages 285–326, https://doi.org/10.2307/2118330

[13] Freeman, Richard B. "Employment and earnings of disadvantaged young men in a labor shortage economy." (1990).

[14] https://www.statistikdatabasen.scb.se/goto/en/ssd/Kommun17g

[15] https://www.statistikdatabasen.scb.se/goto/en/ssd/KV15LJVAKgrreg

[16] Documentation of R^2 score of the *scikit-learn* library: scikit-learn.org/stable/modules/generated/ sklearn.metrics.r2_score.html